On the inferential implications of decreasing weight structures in mixture models

On the inferential implications of decreasing weight structures in mixture models

Pierpaolo De Blasi
University of Torino and Collegio Carlo Alberto, Torino, Italy

Asael Fabian Martínez
Universidad Autónoma Metropolitana, Mexico

Ramsés H. Mena
Universidad Nacional Autónoma de México, Mexico

Igor Prünster
Bocconi University and BIDSA, Milano, Italy

Abstract

Bayesian estimation of nonparametric mixture models strongly relies on available representations of discrete random probability measures. In particular, the order of the mixing weights plays an important role for the identifiability of component-specific parameters which, in turn, affects the convergence properties of posterior samplers. The geometric process mixture model provides a simple alternative to models based on the Dirichlet process that effectively addresses these issues. However, the rate of decay of the mixing weights for this model may be too fast for modeling data with a large number of components. The need for different decay rates arises. Some variants of the geometric process featuring different decay behaviors, while preserving the decreasing structure, are presented and investigated. An asymptotic characterization of the number of distinct values in a sample from the corresponding mixing measure is also given, highlighting the inferential implications of different prior specifications. The analysis is completed by a simulation study in the context of density estimation. It shows that by controlling the decaying rate, the mixture model is able to capture data with a large number of components.

1 Introduction

Mixture models have been subject of interest in many and different contexts due to their flexibility. Their study can be traced back to the work of Pearson [1894]. Extensive accounts can be found, e.g., in Titterington et al. [1985], McLachlan and Peel [2000] and Frühwirth-Schnatter [2006]. A mixture density can be written as

$$f(y) = \sum_{j=1}^{m} w_j \kappa(y; x_j), \tag{1}$$

where $(w_1, \ldots, w_m)$ are the mixture weights, i.e. non-negative values summing up to one, and κ is a density function in y with parameter x_j. According to the employed methodology the number of components, m, in the mixture can be finite or infinite and (w_j, x_j) fixed or random. It is useful to describe the mixture density (1) in terms of an underlying mixing measure, that is

$$f(y) = \int_{\mathbb{X}} \kappa(y; x)\tilde{p}(\mathrm{d}x), \qquad \tilde{p}(\cdot) = \sum_{j=1}^{m} w_j \delta_{x_j}(\cdot), \tag{2}$$

where $\tilde{p}$ defines a random discrete probability measure on $\mathbb{X}$. Here, $(w_j)_{j\geq 1}$ is a sequence of $(0, 1)$-valued random variables summing up to one almost surely (a.s.), and $(x_j)_{j\geq 1}$ is a sequence of independent and identically distributed (iid) random variables from a non-atomic distribution ν_0, and both sequences are taken as independent. Under a Bayesian nonparametric framework, i.e. when $m = \infty$, Lo [1984] studied model (2) when $\tilde{p}$ is given by the Dirichlet process [Ferguson, 1973]. The model, known as the *Dirichlet process mixture* model, gained huge popularity with the work by Escobar and West [1995], where a concrete computational implementation was presented. This model is nowadays widely used in practice thanks also to the availability of effective algorithms.

There are different strategies to provide explicit constructions for the sequence (w_j). Perhaps, one of the most popular approaches is the so-called *stick-breaking* representation, where weights are defined as

$$w_1 = v_1, \qquad w_j = v_j \prod_{l<j} (1 - v_l) \quad j \geq 2, \tag{3}$$

for $(v_j)_{j\geq 1}$ a sequence of independent $(0, 1)$-valued random variables. The only restriction for this construction is that $\sum_{i\geq 1} \log \mathbf{E}(1-v_i) = -\infty$ [Ghosal and van der Vaart, 2017]. The Dirichlet process mixture model is recovered by setting v_j to be iid and beta distributed of parameters $(1, c)$ for some $c > 0$ that corresponds to the total mass parameter of the Dirichlet process. This model constitutes the most representative example of mixture densities in Bayesian nonparametrics, and many other models, with a more complex weight structure, have been conceived based on it; see, e.g. Shi et al. [2019], Quinlan et al. [2018], Nguyen [2010], Griffin and Steel [2011] and Scarpa and Dunson [2014]. Furthermore, their range of application is very wide; Gutiérrez and Quintana [2011] and Wade et al. [2014], to mention just a few, made use of these different models.

A second approach is to make m random in (2), and specify the distribution of $(w_1, \ldots, w_m)$ conditional on m. The resulting model can be expressed as a *mixture of finite mixtures*. A standard choice for the distribution of $(w_1, \ldots, w_m)$ is the symmetric Dirichlet distribution [see, for example Richardson and Green, 1997]. The resulting random discrete probability measure $\tilde{p}$ can be seen to belong to the family of Gibbs-type prior with negative parameter [c.f. De Blasi et al., 2013].

A somewhat different direction in defining a mixture model (2) has been taken by Fuentes-García et al. [2010]. It entails a simple structure for the weights, simpler in that their randomness is determined by a single random parameter. It starts with a finite mixture model with equal weights,

$$f(y|r) = \frac{1}{r} \sum_{i=1}^{r} \kappa(y; x_i). \tag{4}$$

Next, let the number of components r be distributed according to $\pi(\cdot; \theta)$, a probability mass on $\mathbb{Z}^+$ whose parameter θ is random. Then, marginalizing over the distribution of r, one

obtains

$$f(y) = \sum_{r=1}^{\infty} \pi(r;\theta) f(y|r) = \sum_{r=1}^{\infty} \frac{\pi(r;\theta)}{r} \sum_{i=1}^{r} \kappa(y;x_i) = \sum_{j=1}^{\infty} w_j \kappa(y;x_j),$$

which is an infinite mixture density with random weights given by

$$w_j = \sum_{r=j}^{\infty} \frac{\pi(r;\theta)}{r}, \quad j = 1, 2, \ldots. \tag{5}$$

An important characteristic of this construction is that the weights are decreasing a.s., i.e. $w_j > w_{j+1}$, for all $j \geq 1$. Note that those obtained through the stick-breaking representation (3) are not stochastically ordered, rather only in the mean: $\mathbf{E}[w_j] > \mathbf{E}[w_{j+1}]$. Indeed, the distribution of the weights of the Dirichlet process in decreasing order is known as the Poisson–Dirichlet distribution [Kingman, 1993, Section 9.6], however such a representation is not useful for estimation purposes. In the case of mixture of finite mixtures, a similar marginalization argument as in (5) will not yield decreasing weights. To see why, let $m \sim \pi(m)$ in (1) and $(w_{1,m}, \ldots, w_{m,m})$ be the weights in (1) with the dependence in distribution on m made explicit in the notation. For example, take $(w_{1,m}, \ldots, w_{m,m}) \sim \text{Dir}(\gamma, \ldots, \gamma)$, so that $w_{j,m} \sim \text{beta}(\gamma, (m-1)\gamma)$ for $j = 1, \ldots, m$. Marginalization over the distribution of m yields

$$f(y) = \sum_{j=1}^{m} w'_j \kappa(y;x_j), \qquad w'_j = \sum_{m=j}^{\infty} w_{j,m} \pi(m), \quad j = 1, 2, \ldots,$$

where, however, the equality in distribution $w'_{j+1} \stackrel{d}{=} w'_j - w_{j,j}\pi(j)$ does not imply $w'_j > w'_{j+1}$ a.s. anymore.

Going back to model (5), for the specific case where r follows a negative binomial distribution with integer parameter $s = 2$, the weights are found to have a closed form given by

$$w_j = p(1-p)^{j-1}, \quad j = 1, 2, \ldots,$$

with $p \in (0, 1)$. See Section 2 for details. The resulting random probability measure $\tilde{p}$, known as the *geometric process*, is obtained by having $p = w_1$ random with some prior distribution. It is important for this prior distribution to be supported on $(0, 1)$ so that, in particular, $P(w_1 < \epsilon) > 0$ for every $\epsilon > 0$, since the latter has been shown to guarantee that the random probability measure $\tilde{p}$ has full support on the space of distributions over the reference space $\mathbb{X}$, cf. Corollary 3 in [Bissiri and Ongaro, 2014]. The corresponding mixture model represents a simpler yet appealing alternative to models like those based on the stick-breaking representation. Indeed, the geometric process has been successfully applied in different problems; besides density estimation [Fuentes-García et al., 2010], it has been used in regression [Fuentes-García et al., 2009], dependent models [Mena et al., 2011, Hatjispyros et al., 2018], classification [Gutiérrez et al., 2014], and others. Furthermore, having the weights decreasingly ordered alleviates the label switching effect in the implementation of the posterior sampler and allows for a better interpretation of the weight of each component within the whole sample, thus improving the identifiability of the mixture model [Mena and Walker, 2015]. However, the rate of decay of the weights $(w_j)_{j \geq 1}$ can be too rapid for modeling data with a large number of components. In fact, in these cases an accurate reconstruction of the data generating density requires the mixture to feature far too many

components with small weights, which affects the convergence of the posterior sampler. Hence, it is important to derive alternative models with a more flexible rate of decay.

In this paper, we explore some other sequences of mixing weights derived from construction (5). In Section 2, we present two specific cases, one of them can be thought as a generalization of the geometric process. A discussion on the asymptotic behavior for the expected number of groups, as the sample size increases, in a species sampling model setting is presented in Section 3. This illustrates the effect of the decaying rate of the weights when a multinomial sampling from the mixing distribution is considered. Section 4 provides a Markov Chain Monte Carlo (MCMC) algorithm (with further details given in the Appendix), and in Section 5, we perform a simple yet effective simulation study in the context of density estimation with univariate as well multivariate data. This study showcases how to tune the decreasing weights structure to reconstruct the data generating density. Some concluding remarks are provided in Section 6.

2 Extensions of the geometric process

The construction in (5) relies on the choice of the distribution $\pi(\cdot;\theta)$ of r. As mentioned in the Introduction, the geometric process corresponds to a particular instance of $\pi(\cdot;\theta)$. Here we provide details of the latter derivation and explore other choices of $\pi(\cdot;\theta)$ that yield different decay behaviors. The distribution of the weights is determined by the prior distribution on the weight parameter θ. In particular, as mentioned above, it is important that the prior on θ yields $P(w_1 < \epsilon) > 0$ for every $\epsilon > 0$ as this guarantees the full support of $\tilde{p}$.

We consider two different specifications for the distribution $\pi(\cdot;\theta)$ of r: the negative binomial and the Poisson distributions. In both cases, we modify the support by excluding the value zero, in order to apply construction (5). As for the former, let a random variable X follow a shifted negative binomial distribution, $X \sim \mathrm{NB}_1(s,p)$, with parameters (s,p), for $s > 0$ and $0 < p < 1$, if its probability distribution is given by

$$\pi(x;s,p) = \binom{x+s-2}{x-1} p^s (1-p)^{x-1}, \quad x = 1,2,\ldots. \tag{6}$$

Furthermore, $\mathbf{E}(X) = 1 + (1-p)s/p$, and $\mathrm{Var}(X) = (1-p)s/p^2$. Substituting $\pi(\cdot;s,p)$ in (5), we obtain

$$w_j = \frac{1}{j}\binom{j+s-2}{j-1} p^s (1-p)^{j-1} {}_2F_1(j+s-1,1,j+1;1-p), \tag{7}$$

for $j = 1,2,\ldots,$ where

$${}_2F_1(a,b,c;z) = \sum_{k=0}^{\infty} \frac{(a)_{k\uparrow}(b)_{k\uparrow}}{(c)_{k\uparrow}} \frac{z^k}{k!},$$

is the Gaussian hypergeometric function, and $(x)_{n\uparrow}$ denotes the Pochhammer symbol defined as $(x)_{n\uparrow} = x(x+1)\cdots(x+n-1)$ with the convention $(x)_{0\uparrow} = 1$. The formula above simplifies when s is an integer. By direct calculation involving the geometric series, $s = 2$ leads to

$$\begin{aligned} w_j &= \sum_{r \geq j} \frac{1}{r} r p^2 (1-p)^{r-1} = p^2 (1-p)^{j-1} \sum_{i \geq 0} (1-p)^i \\ &= p^2 (1-p)^{j-1} \frac{1}{p} = p(1-p)^{j-1}, \quad j = 1,2,\ldots, \end{aligned}$$

that is the geometric distribution. An explicit form is also obtained for $s = 3$ by exploiting the derivative of the geometric series to get

$$w_j = p(1-p)^{j-1}\frac{1+jp}{2}, \quad j = 1, 2, \ldots. \tag{8}$$

Similar formulae can be obtained for $s = 4, 5, \ldots$ by taking further derivatives. We set $\theta = p$ and endow it with a prior distribution. In all cases, when p is random with distribution supported on $(0, 1)$, w_1 is is supported on $(0, 1)$ too, hence $\tilde{p}$ has full support.

The second case is based on the Poisson distribution. Performing the same shift as before, we say that a random variable X follows a shifted Poisson distribution, $X \sim \text{Poi}_1(\lambda)$, with parameter $\lambda > 0$, if its probability distribution is such that

$$\pi(x; \lambda) = \frac{\lambda^{x-1}e^{-\lambda}}{(x-1)!}, \quad x = 1, 2, \ldots. \tag{9}$$

Moreover, $\mathbf{E}(X) = 1 + \lambda$ and $\text{Var}(X) = \lambda$. In this case, setting $\theta = \lambda$ in (5), the corresponding weights become

$$w_j = \frac{\Gamma(j) - \Gamma(j, \lambda)}{\lambda\Gamma(j)}, \tag{10}$$

for $j = 1, 2, \ldots$, where $\Gamma(a, z)$ is the (upper) incomplete Gamma function, which has the following representation, for any positive integer a,

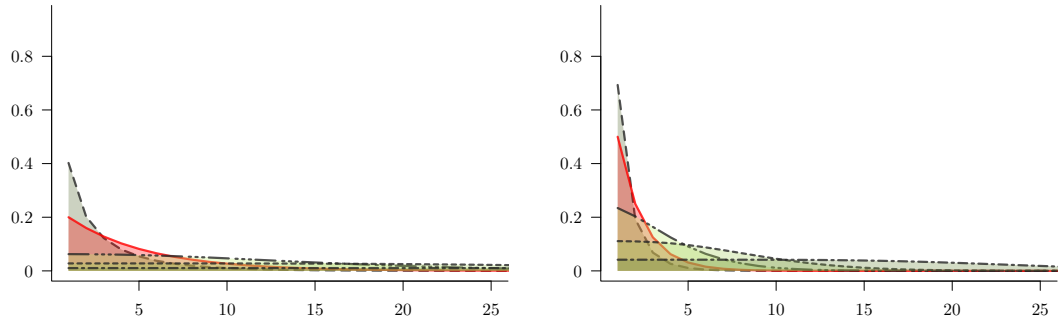
$$\Gamma(a, z) = \frac{\Gamma(a)}{e^z}\sum_{k=0}^{a-1}\frac{z^k}{k!}.$$

Note that, having λ supported on $\mathbb{R}_+$ implies that w_1 is supported on $(0, 1)$, as $\lim_{\lambda\downarrow 0}\Gamma(j, \lambda) = \Gamma(j)$ and $\lim_{\lambda\to\infty}\Gamma(j, \lambda) = 0$, and the full support of $\tilde{p}$ follows.

For these cases the decay rate varies according to the value of their corresponding weight parameters θ. In Figure 1, different examples for these two cases are displayed. A convenient way of characterizing the tail behavior of the weights is in terms of the asymptotic distribution of the random variable denoting the number of distinct values K_n observed in a n-sample from $\tilde{p}$. In the next section, we use Karlin's theory to show how heavy tailed weights correspond to more new values as n increases, i.e. to a faster rate of increase in K_n.

3 Asymptotic behavior of $\mathbf{E}(K_n)$ in species sampling problems

An key characteristic to investigate for discrete random probability measures is the distribution of the (random) number K_n of distinct values in a sample from the probability measure and, in particular, how it increases as the sample size, n, increases. Understanding such a behavior is crucial for effective modeling with discrete random probability measures. See, e.g., [Lijoi et al., 2007a,b, Lijoi, Antonio and Nipoti, Bernardo and Prünster, Igor, 2014, De Blasi et al., 2015]. Here we discuss the asymptotic behavior of K_n in a species sampling problem context by focusing on the geometric process and its negative binomial extensions discussed in Section 2. The following notation is adopted throughout the section: for two sequences $(a_n), (b_n)$, $a_n = O(b_n)$ means that $|a_n| \leq C|b_n|$ for all sufficiently large n and for some positive constant C independent of n; $a_n \sim b_n$ means that $a_n/b_n \to 1$ as $n \to \infty$; when $(a_n), (b_n)$ are random, $a_n \sim_{a.s.} b_n$ means that $a_n/b_n \to 1$ almost surely.



(a) NB weights for $p = 0.2$ and different values of s: 1, 2, 5, 10 and 25.

(b) NB weights for $p = 0.5$ and different values of s: 1, 2, 5, 10 and 25.

(c) NB weights for $p = 0.8$ and different values of s: 1, 2, 5, 10 and 25.

(d) Poisson weights for different values of λ: 0.1, 1, 5, 10 and 20.

Figure 1: Illustration of the different decay rates for the negative binomial (NB) and Poisson constructions of the mixing weights. For the NB case, the geometric process ($s = 2$) is displayed in red. Points are connected by straight lines for visual simplification.

Let $w_j = p(1-p)^{j-1}$, $j = 1, 2, \ldots$, $p \in [0, 1]$, be the geometric probabilities and $\phi(p)$ be the prior density on the success probability of the geometric distribution. Accordingly, here K_n denotes the number of distinct values generated by n draws from the geometric process. The distribution of K_n is determined by the process of random sampling and the randomness of the (w_j). In order to study $\mathbf{E}(K_n)$, we exploit the law of total expectation by considering first the so called occupancy problem in the case of fixed weights. To this aim, the key quantity to consider is the number of geometric probabilities not smaller than $x \in (0, 1)$, which is found to be

$$\overrightarrow{\nu}(x, p) := \#\{j : \ p(1-p)^{j-1} \geq x\} = \left\lfloor \frac{\log(x/p)}{\log(1-p)} + 1 \right\rfloor \mathbf{1}_{(p \geq x)}.$$

Here $\lfloor x \rfloor$ is the integer part of x and $\mathbf{1}_A(\cdot)$ is the indicator function of set A. Let $\mathbf{E}(K_n|p)$ be the number of distinct values in an iid sample of size n from the geometric distribution of parameter p. According to Karlin [1967], $\mathbf{E}(K_n|p) \sim_{a.s.} \overrightarrow{\nu}(\frac{1}{n}, p)$ as $n \to \infty$, so by the law of total expectation and Fubini theorem,

$$\mathbf{E}(K_n) \sim I(\tfrac{1}{n}), \quad \text{where} \quad I(x) = \int_0^1 \overrightarrow{\nu}(x, p)\phi(p)\mathrm{d}p. \tag{11}$$

Hence the asymptotic behavior of $\mathbf{E}(K_n)$ depends on the limiting behavior of the integral $I(x)$ for $x \downarrow 0$. For illustration, let $\phi(p)$ be the uniform distribution on $(0, 1)$. In order to

study the limiting behavior of $I(x)$ as $x \downarrow 0$, by a change of variable, it is sufficient to focus on

$$J(x) = \int_0^{\log 1/x} \left(\log \frac{1}{x} - t \right) \mathrm{d}f(t), \quad \text{where} \quad \frac{\mathrm{d}f(t)}{\mathrm{d}t} = \frac{\mathrm{e}^{-t}}{-\log(1-\mathrm{e}^{-t})},$$

as $J(x) \leq I(x) \leq 1 - x + J(x)$. The integral $J(x)$ corresponds to the Riemann-Liouville integral, or fractional integral, ${}_1f(\cdot)$ of $f(t)$, according to the definition

$${}_\alpha f(x) = \frac{1}{\Gamma(\alpha+1)} \int_0^x (x-t)^\alpha \mathrm{d}f(t),$$

evaluated at $\log 1/x$. According to Bingham et al. [1987, Page 58], for $f(t)$ non decreasing with $f(0) = 0$, f is regularly varying at infinity with exponent β iff ${}_\alpha f(x)$ is regularly varying at infinity with exponent $\alpha + \beta$ and each implies

$$\frac{{}_\alpha f(x)}{x^\alpha f(x)} \to \frac{\Gamma(\beta+1)}{\Gamma(\alpha+\beta+1)}.$$

In the present case, $\mathrm{d}f(t)/\mathrm{d}t$ behaves like a distribution function on $\mathbb{R}_+$, so we have $\beta = 1$. It follows that ${}_1f(\cdot)$ is regularly varying at infinity with exponent $\alpha+\beta = 2$ and the constant of proportionality is $\Gamma(\beta+1)/\Gamma(\alpha+\beta+1) = 1/2$, so that $J(x) \sim (\log 1/x)^2/2$ as $x \to 0$ and we conclude that

$$\mathbf{E}(K_n) \sim \frac{1}{2}\big(\log n\big)^2,$$

in contrast to $\mathbf{E}(K_n|p) \sim \frac{1}{|\log(1-p)|} \log n$ in the fixed (w_j) case. Faster (slower) rates are obtained by changing the prior $\phi(p)$ so to shift mass to lower (higher) values of p. It can be shown that, for $p \stackrel{d}{=} \mathrm{e}^{-X}$ for $X \sim \text{gamma}(a, 1)$, $a > 0$,

$$\mathbf{E}(K_n) \sim \frac{1}{a(a+1)} (\log n)^{a+1}.$$

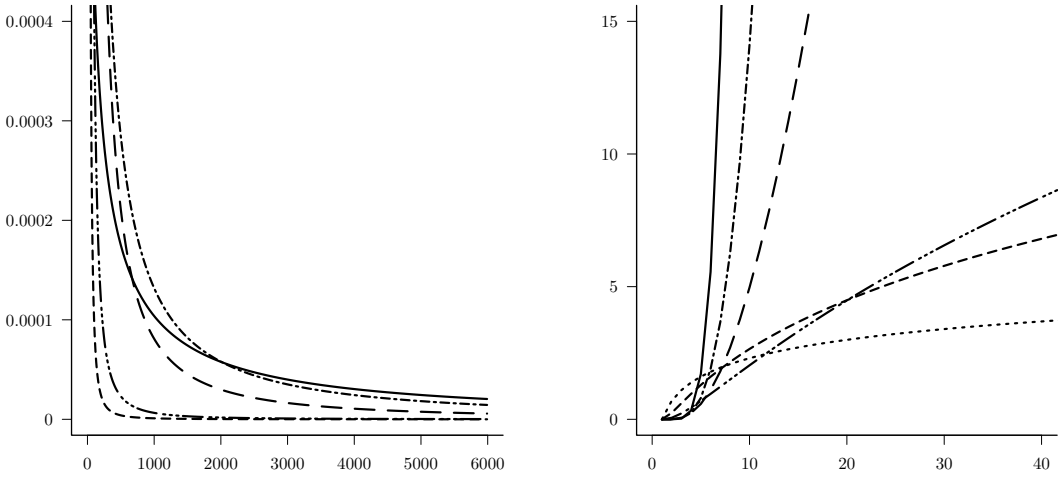
For the Dirichlet process case with total mass parameter $c > 0$, the asymptotic behavior of $\mathbf{E}(K_n)$ is known to be proportional to $c \log n$ [Korwar and Hollander, 1973]. In our case, by tuning the prior $\phi(p)$, a whole range of logarithmic behaviors can be achieved. See Figure 2 for an illustration. Other examples of growth rates different than $\log n$ can be found in Camerlenghi et al. [2019] and Gnedin et al. [2006].

Recall that the geometric process corresponds to construction (5) when the distribution of r is shifted negative binomial with $s = 2$, cf. (6). Next we investigate next the asymptotic behavior of $\mathbf{E}(K_n)$ for s larger than 2. A comparison with the geometric process case is meaningful only if made for a fixed prior $\phi(p)$, so for illustrative purposes we stick to the uniform prior here. As mentioned in the previous section, the weights (w_j) admit an explicit form for an integer s, as the tools developed above, based on the quantity $\overrightarrow{\nu}(x, p)$, can be carried out. We report here a result for the case $s = 3$, cf. (8), but similar results are believed to hold for s larger than 3. It says that the leading term in the asymptotic expansion of $\mathbf{E}(K_n)$ is not affected by the value of s, specifically,

$$\mathbf{E}(K_n) - \frac{1}{2}\big(\log n\big)^2 = O(\log n \, \log\log n).$$

Performing an analogous study for the Poisson case is difficult with the present techniques due to the formula of the weights (w_j). The asymptotic evaluation of $\mathbf{E}(K_n)$ is still of type (11), namely

$$\mathbf{E}(K_n) \sim \int_0^1 \overrightarrow{\nu}(1/n, \lambda)\phi(\lambda)\mathrm{d}\lambda, \quad \overrightarrow{\nu}(x, \lambda) = \#\left\{ j : \ \tfrac{\Gamma(j)-\Gamma(j,\lambda)}{\lambda\Gamma(j)} \geq x \right\},$$



(a) Geometric weights

(b) Asymptotic behavior of $\mathbf{E}(K_n)$. The case $\log(n)$ is displayed with a dotted line.

Figure 2: Geometric weights and asymptotic behavior of $\mathbf{E}(K_n)$ where the prior for p is given by e^{-X} with $X \sim \text{gamma}(a, 1)$. Different values for a were used: **– –** 1, **–··–** 2, **——** 5, **–·–·–** 7 and **———** 10. Points are connected by straight lines for visual simplification.

where $\phi(\lambda)$ is the prior on the mean parameter λ of the Poisson distribution and $\overrightarrow{\nu}(x, \lambda)$ is the number of weights w_j not smaller than x according to (10). Unfortunately, the lack of an explicit form of w_j makes the evaluation of the integral in the display above hard. We will further research this direction elsewhere.

4 Sampling scheme for density estimation

In this section, we provide an MCMC algorithm to draw samples from the posterior distributions of interest. A mixture model of the form of (4) can be written hierarchically as

$$\begin{aligned} y_i|x, r_i, d_i &\overset{\text{ind}}{\sim} \kappa(y_i; x_{d_i}), \qquad i = 1, \ldots, n \\ d_i|r_i &\overset{\text{ind}}{\sim} \text{U}(d_i; 1, r_i) \\ r_i|\theta &\overset{\text{iid}}{\sim} \pi(r_i; \theta, \psi) \\ x_j &\overset{\text{iid}}{\sim} \nu_0(x_j; \xi), \qquad j = 1, \ldots, m \\ \theta &\sim \phi(\theta; \omega), \end{aligned}$$

for $m := \max(r_1, \ldots, r_n)$, and where $\text{U}(\cdot; 1, r)$ is the uniform distribution on the set of integers $\{1, 2, \ldots, r\}$, ν_0 is a non-atomic distribution (identified by its corresponding density function) and ψ, ξ and ω are known finite dimensional parameters. Variables $(d_1, \ldots, d_n)$ are membership variables associating each observation y_i to the d_ith mixture component. The distributions π and ϕ are chosen according to the negative binomial or Poisson case. Therefore, we can implement a Gibbs sampler whose full conditional distributions are the following:

1. the full conditional for the kernel parameter is

$$p(x_j|\cdots) \propto \nu_0(x_j;\xi) \prod_{d_i=j} \kappa(y_i;x_j),$$

for $j = 1,\ldots,m$.

2. For the membership variables, we have

$$p(d_i|\cdots) \propto \kappa(y_i;x_{d_i})\mathbf{1}(1 \le d_i \le r_i),$$

for $i = 1,\ldots,n$.

3. For the integer-valued r_i, the full conditional is given by

$$p(r_i|\cdots) \propto \frac{\pi(r_i;\theta,\psi)}{r_i}\mathbf{1}(r_i \ge d_i).$$

4. Finally, parameter θ is updated by

$$p(\theta|\cdots) \propto \phi(\theta;\omega)\prod_{i=1}^{n}\pi(r_i;\theta,\psi).$$

The last two steps are different for the negative binomial and the Poisson case, as detailed next.

Negative binomial case Assuming $r_i \sim \mathrm{NB}_1(s,p)$ for $i = 1,\ldots,n$, so that $\theta = p$ and $\psi = s$, and $p \sim \mathrm{beta}(\alpha,\beta)$, the conditional distributions are

$$p(r_i|\cdots) \propto \binom{r_i+s-2}{r_i}(1-p)^{r_i}\mathbf{1}(r_i \ge d_i),$$

that is, a negative binomial distribution with parameters $(s-1,p)$ truncated at d_i, and

$$p(p|\cdots) \propto p^{sn+\alpha-1}(1-p)^{\sum_{i=1}^n r_i+\beta-n-1},$$

which is a beta distribution with parameters $(sn+\alpha, \sum_{i=1}^n r_i+\beta-n)$. A simulation procedure for the truncated distribution is presented in A.

Poisson case If it is assumed $r_i \sim \mathrm{Poi}_1(\lambda)$ for all $i = 1,\ldots,n$, so that $\theta = \lambda$ (no ψ needed here), and $\lambda \sim \mathrm{gamma}(\gamma,\delta)$, the corresponding conditional distributions are

$$p(r_i|\cdots) \propto \frac{\lambda^{r_i}}{r_i!}\mathbf{1}(r_i \ge d_i),$$

which is a Poisson distribution with parameter λ truncated at d_i, and

$$p(\lambda|\cdots) \propto \lambda^{\sum_{i=1}^n r_i+\gamma-n-1}e^{-(\delta+n)\lambda},$$

a gamma distribution with parameters $(\sum_{i=1}^n r_i+\gamma-n, \delta+n)$. In A, a procedure to simulate from the truncated distribution is also described.

Regarding the density estimation, we use the following Monte Carlo estimator

$$\hat{f}(y) = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{m^{(t)}} w_j^{(t)}\kappa(y;x_j^{(t)}), \tag{12}$$

where the mixing weights are computed using either (7) (or explicit formulae like (8) for integer-valued s) together with the sampled $p^{(t)}$ or (10) together with the sampled $\lambda^{(t)}$, and $m^{(t)} = \max(r_1^{(t)}, \ldots, r_n^{(t)})$ is the number of kernel parameters $x_j^{(t)}$ sampled at iteration t.

As mentioned in the Introduction, having a mixing measure $\tilde{p}$ with heavy tailed weights is expected to provide a better density estimate with data featuring a large group structure. In the next section, we explore the performance of the models under study with both univariate and multivariate simulated data in this setting.

5 Simulation study

5.1 Univariate case

We use a Gaussian kernel with unknown mean and variance, so $\kappa(y; x_j) = N(y; \mu_j, 1/\tau_j)$. A normal–gamma prior with parameters (m, c, a, b) is used for the parameters (μ_j, τ_j), that is

$$p(\mu_j, \tau_j) = \mathrm{N}(\mu_j; m, c/\tau_j)\mathrm{Ga}(\tau_j; a, b).$$

Thus, the posterior distribution is conjugate with parameters (m', c', a', b') given by

$$m' = \frac{cn_j\bar{y}_j + m}{cn_j + 1}, \qquad a' = \frac{n_j}{2} + a,$$
$$c' = \frac{c}{(cn_j + 1)}, \qquad b' = \frac{n_j(\bar{y}_j - m)^2}{2(cn_j + 1)} + \frac{S_j}{2} + b,$$

where $\bar{y}_j = \frac{1}{n_j}\sum_{d_i=j} y_i$, $S_j = \sum_{d_i=j}(y_i - \bar{y}_j)^2$ and $n_j = \sum_{j=1}^{n} \mathbf{1}(d_i = j)$.

Given this model, we test the two cases under different scenarios, i.e. using two datasets. The first dataset is a sample of size 1 000 from the following mixture of 5 components:

$$f(y) = \tfrac{39}{100}\mathrm{N}(y; 3, 1) + \tfrac{21}{100}\mathrm{N}(y; 5.5, 4) + \tfrac{15}{100}\mathrm{N}(y; 15, 1) + \tfrac{10}{100}\mathrm{N}(y; 20, 0.25) + \tfrac{15}{100}\mathrm{N}(y; 22, 16). \tag{13}$$

The second dataset consists also on a sample of size 1 000, but from a mixture with 10 components:

$$f(y) = \sum_{j=1}^{10} \frac{1}{10}\mathrm{N}(y; 6j - 33, 1). \tag{14}$$

For both datasets, 50 000 iterations of the MCMC scheme were obtained; the first 30 000 of them were discarded. We assign vague priors by setting $(m, c, a, b) = (0, 1000, 0.01, 0.01)$ for the kernel hyperparameters, and letting $p \sim \mathrm{beta}(1, 1)$, i.e. uniformly distributed on $(0, 1)$, for the negative binomial case, and $\lambda \sim \mathrm{gamma}(0.01, 0.01)$ for the Poisson case. Additionally, we fix the second parameter for the negative binomial case, s, taking the values 1, 2 (corresponding to the geometric process), 5, 10 and 25.

Figures 3 and 4 display the estimated density for each dataset and each scenario. Furthermore, for the case with a large number of components, model (14), Figure 5 displays the estimated and true mixing weights (panel 5a) and the estimated posterior distribution of K_n (panel 5b). For the sake of comparison, we also include the density estimator obtained from the Dirichlet process mixture model with a fixed total mass parameter c. This was implemented via a slice sampler [Kalli et al., 2011], and we use the same MCMC specifications, and prior settings; additionally, the total mass parameter c was fixed in such a way that $\mathbf{E}(K_n)$ matches the true number of components.

Examining the results for the first dataset, drawn from model (13), the posterior density estimations for the data with a relatively small number of components are all similar (Figure 3). However, for the second dataset featuring a larger number of components (model 14), the density estimates show more evident differences (Figure 4). Indeed, in order to capture the density that features a relatively large number of similar modes, a flexible weight structure plays a crucial role. One can appreciate that, for the scenarios where the posterior mixing weights have a heavy tailed weights decay (Figure 5a), the estimated weight components replicate better those of the data generating process and the posterior distribution of K_n moves towards the true number of components as the weight tail gets heavier (Figure 5b). The Dirichlet model instead underestimates K_n while failing to reconstruct three of the ten modes of the data generating density, cf. Figure 4. This is due to the mixing weights decreasing too fast, cf. Figure 5a, so the posterior mixing measure is unable to place enough mass at each of the ten modes.

Here, it is worth emphasizing that these results get closer to Feller [1943] desirable interpretation, of the estimated mixing weights of a mixture model representing the true proportions, a feature that is rarely recovered with other mixture model approaches. For the negative binomial case, this effect is achieved for big values of s; actually, it can be appreciated that as it increases, the density estimator adjusts better to the original function. The above observation clearly ponders over the flexibility of the model to adapt a small or large number of components. For instance, in the Poisson case, where the bigger the λ, the heavier-tailed weights decay (cf. Figure 1d), placing a prior on λ allows the model to adapt to the required tail behavior. Similar adaptation follows as one varies the value for s. Indeed, if we let $p = 1 - \frac{\lambda}{s+\lambda}$, for some $\lambda, s > 0$, then $0 < p < 1$. Hence, if we set (p, s), the parameters of the NB_1 probability function (6), we have

$$\binom{x+s-2}{x-1} p^s (1-p)^{x-1} = \frac{\lambda^{x-1}}{(x-1)!} \frac{\Gamma(x+s-1)}{(s+\lambda)^{x-1}\Gamma(s)} \left(1+\frac{\lambda}{s}\right)^{-s}.$$

Taking its limit as $s \to \infty$, the middle term converges to one, whereas for the last term, we have $\left(1+\frac{\lambda}{s}\right)^{-s} \to e^{-\lambda}$. Thus, the probability function (9), i.e. the shifted Poisson distribution Poi_1, is recovered. This tells us that when randomizing the parameters we can achieve similar adaptation in the tails with either model, negative binomial or Poisson.

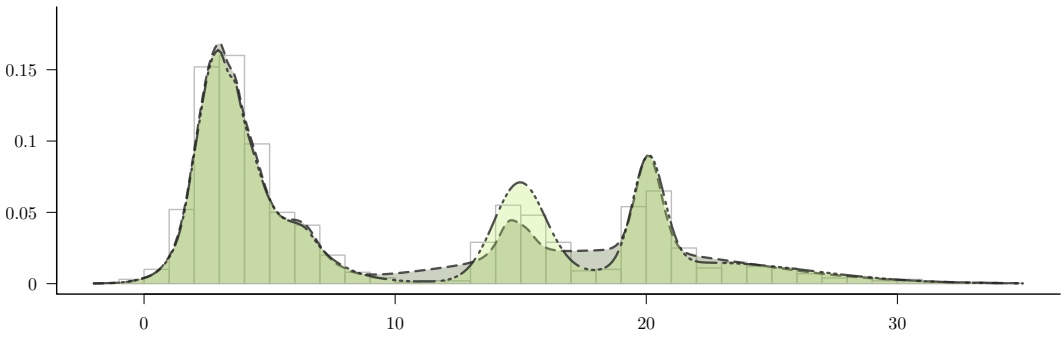
5.2 Multivariate case

We now explore the performances of the negative binomial mixture models in a multivariate setting in comparison with the Dirichlet model. A bivariate Gaussian kernel $\kappa(y; x_j) = N(y; \mu_j, \Sigma_j)$ is used, so $x_j = (\mu_j, \Sigma_j)$, where the mean vector μ_j and the covariance matrix Σ_j are both unknown, and a conjugate normal-inverse Wishart prior with parameters (m, Λ, v) is chosen, that is

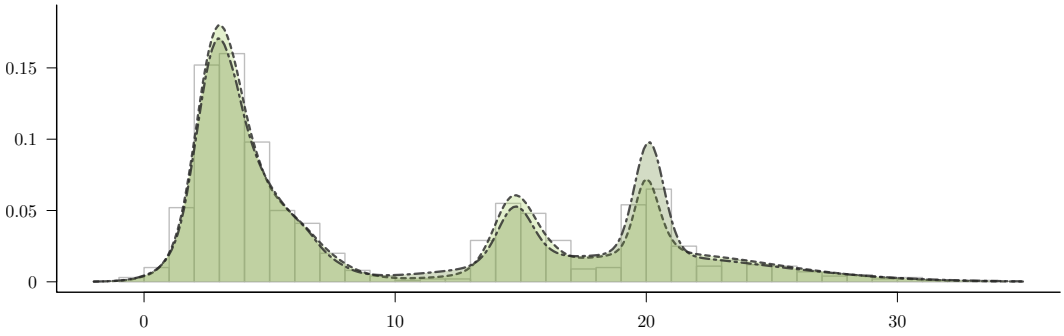
$$p(\mu_j, \Sigma_j) = N(\mu_j, m, \Sigma_j) iW(\Lambda^{-1}, v).$$

Thus, the corresponding posterior distribution, a normal-inverse Wishart, has parameters (m', Λ', v') given by

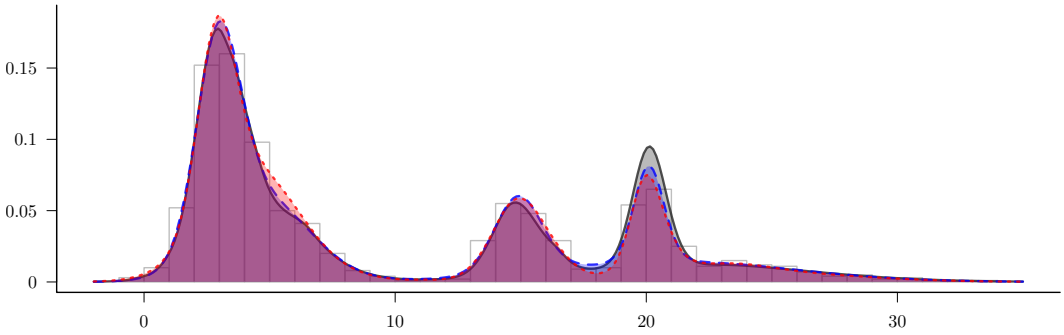
$$m' = \frac{m + n_j \bar{y}_j}{n_j + 1}, \qquad \Lambda' = \Lambda + S_j + \frac{n_j}{n_j + 1}(\bar{y}_j - m)(\bar{y}_j - m)^T,$$
$$v' = v + n_j,$$



(a) Estimated density for the NB case with $s = 1$ – –, and $s = 5$ – · –.



(b) Estimated density for the NB case with $s = 10$ · · · ·, and $s = 25$ – · –.



(c) Estimated density for the Poisson case (black), geometric, $s = 2$, (red/short dash) and Dirichlet (blue/long dash) processes.

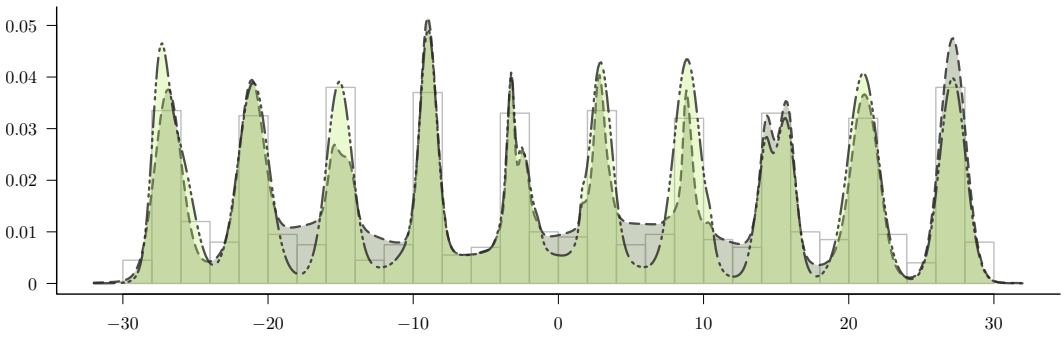
Figure 3: Estimated density of the dataset drawn from model (13) for each prior specification.

where $\bar{y}_j = \frac{1}{n_j}\sum_{d_i=j} y_i$, $S_j = \sum_{d_i=j}(y_i - \bar{y}_j)(y_i - \bar{y}_j)^T$ and $n_j = \sum_{j=1}^{n} \mathbf{1}(d_i = j)$. Convergence of the MCMC sampler is monitored by computing the Hellinger distance. Recall that the Hellinger distance, H, between two densities f and g is defined as

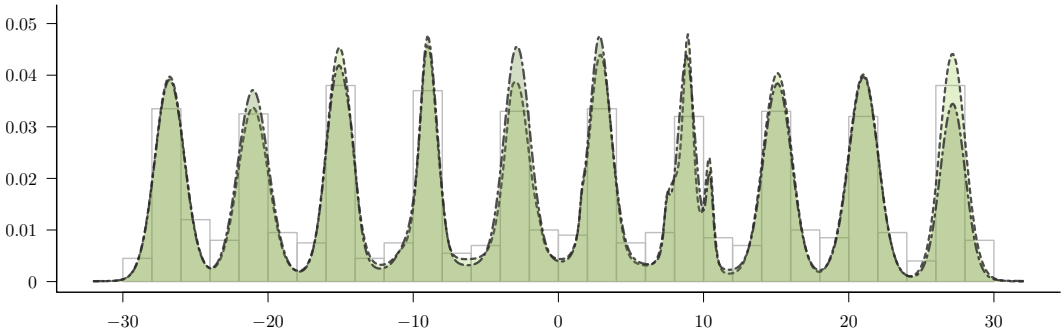
$$H^2(f,g) = \frac{1}{2}\int \left(\sqrt{f(y)} - \sqrt{g(y)}\right)^2 \mathrm{d}y = 1 - \int \sqrt{f(y)g(y)}\mathrm{d}y.$$

Thus, at each MCMC iteration, this distance is computed taking f to be the data generating density and $g := \hat{f}$, the estimated density in Equation (12).

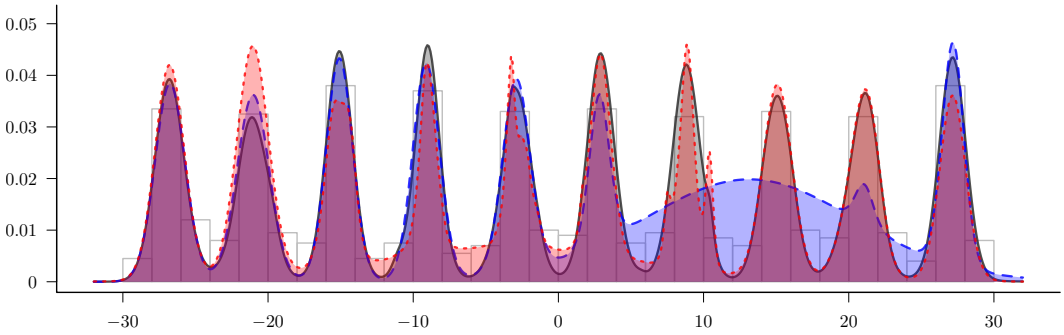
Similarly to the previous illustration (model 14), a sample of size 900 is drawn from the



(a) Estimated density for the NB case with $s = 1$ – –, and $s = 5$ –··–.



(b) Estimated density for the NB case with $s = 10$ ····, and $s = 25$ –··–.



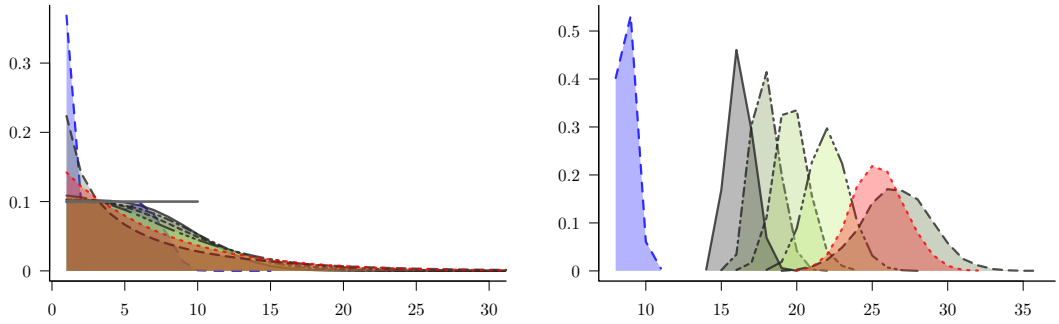
(c) Estimated density for the Poisson case (black), geometric, $s = 2$, (red/short dash) and Dirichlet (blue/long dash) processes.

Figure 4: Estimated density of model (14) for each prior specification.

mixture model

$$f(y) = \frac{1}{9} \sum_{\mu_1 \in M} \sum_{\mu_2 \in M} N((\mu_1, \mu_2), I), \tag{15}$$

for $M = \{-6, 0, 6\}$ and I the identity matrix. Regarding the hyperparameters that determine the distribution of the weights, α, β for the negative binomial and the total mass c for the Dirichlet, we fix them so that $E(K_n) \in \{3, 9, 100\}$ *a priori*. This is accomplished by resorting to the asymptotic evaluations laid out in Section 3 for the negative binomial. Specifically, we compute the integrals as in (11) by Monte Carlo averaging so to tune the parameters of the prior on the parameter p. In the Dirichlet case, the explicit formula



(a) Mixing weights, w_j. The true mixing weights, $w_j = 0.1$ for $j = 1, \ldots, 10$, are displayed in gray.

(b) Distribution of the number of components, K_n.

Figure 5: Posterior results for each prior specification of model (14): negative binomial case for different values of s: – – 1, –·– 5, ···· 10 and –··– 25, and Poisson case (solid black). The geometric, $s = 2$, (red/short dash) and Dirichlet process (blue/long dash) are also displayed. Points are connected by straight lines for visual simplification.

$\mathbf{E}(K_n) = \sum_{i=1}^{n} c/(c+i-1)$ has been used instead. The kernel hyperparameters are

$$m = (-0.0485, -0.0103), \quad \Lambda = \begin{pmatrix} 25.3187 & 0.2742 \\ 0.2742 & 25.2321 \end{pmatrix}, \quad \text{and} \quad v = 4.$$

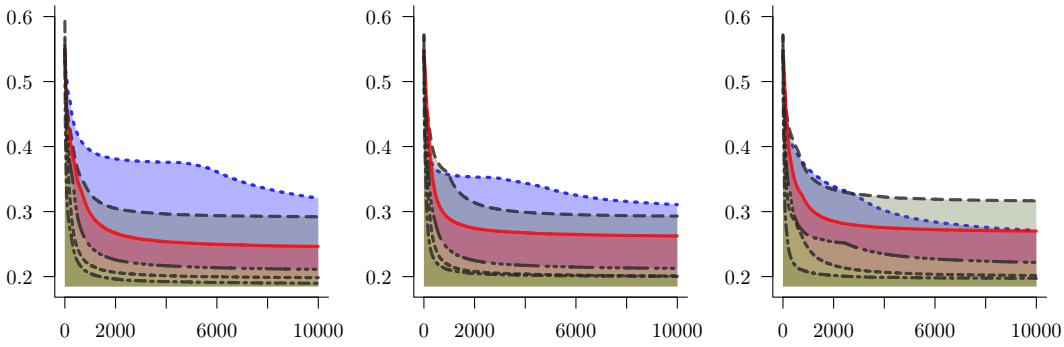
First values are the sample mean and sample covariance matrix for m and Λ, respectively; the value for v allows to have a flat density.

Figure 6 shows the computed distances at each MCMC iteration. The effect of s is apparent: as it increases, the distance decreases faster. This effect is also detectable in the estimated density. The estimated densities for the case $E(K_n) = 3$, where the smallest distances were obtained, are depicted in Figure 7. These were obtained using 2 000 iterations after a burn-in of 8 000. The Dirichlet model is outperformed by all other models in overall accuracy and convergence of the MCMC sampler when $\mathbf{E}(K_n)$ is set to 3 or 9, and by the negative binomial model with $s = 5, 10, 25$ when $\mathbf{E}(K_n) = 100$.

6 Concluding remarks

Decreasing-weight mixture models represent an appealing alternative to other infinite-mixture counterparts like those based on the stick-breaking representation. The simulation studies show that a flexible structure in the decreasing mixing weights improves the accuracy of posterior estimates and enhance the convergence of the posterior sampler, in particular when the data feature a large component structure. Such flexibility can be achieved by widening the choices of the prior on r in the construction of the weights (5) or by selecting a suitable prior for the hyperparameter p in the geometric process. An important finding is that slowly decreasing weights are required in cases where the underlying mixture density is made of many components.

The improvement in the convergence of the posterior sampler is related to the gain in identifiability of the mixture when the weights are decreasing. This is explained by a reduction of the region of the parameter space that the sampler has to explore. Out of



(a) Case $E(K_n) = 3$ *a priori* (b) Case $E(K_n) = 9$ *a priori* (c) Case $E(K_n) = 100$ *a priori*

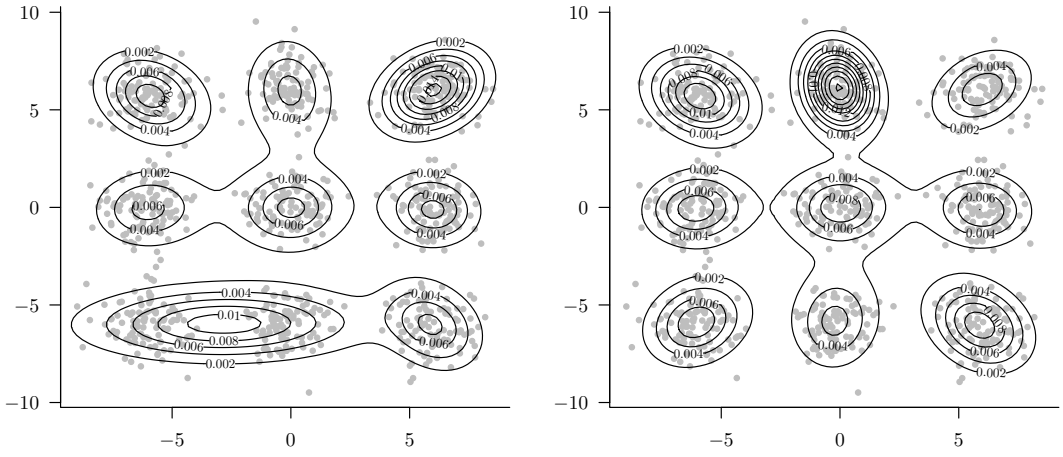
Figure 6: Hellinger distance, computed by iteration, for each MCMC specification of model (15) using negative binomial weights. Different values of s were used: – – 1, – - – 5, ···· 10 and – ·· – 25. The geometric, $s = 2$, (solid/red) and Dirichlet (blue/dotted) processes are also is displayed.

the many alternatives to the Dirichlet process, very few can be implemented via marginal samplers, that is a sampler that integrates out the mixing weights (w_i). When it comes to conditional samplers, it is well known that they suffer from the so called label switching effect, which consists of the posterior of (w_i, x_i) exhibiting multiple modes. This is due to the fact that the augmentation of (w_j) in the conditional methods makes the infinite mixture weakly identifiable since there is a non-null probability of both $w_i > w_j$ and $w_i < w_j$ for i, j close to each other. Cf. the discussion in Papaspiliopoulos and Roberts [2008]. The presence of multiple modes in the posterior of w_i deteriorates the mixing of the sampler since it has to explore different and possibly far away regions of the parameter space. This is typically the case for mixing measures with stick-breaking representation, where the weights are ordered only in the expected value.

Another way of looking at this problem is in terms of the presence of gaps in the labeling of (w_i, x_i) across iterations of the sampler, cf. Mena and Walker [2015]. As explained in Fuentes-García et al. [2010], the original idea behind the geometric process was to alleviate this problem, by creating a model without such gap structure. However the decay of the weights in the geometric process might be too fast to accommodate datasets with many components. Thus there is a concrete need for models, which exhibit decreasing weights as the geometric process, but a slower rate of decay like the ones explored in this paper.

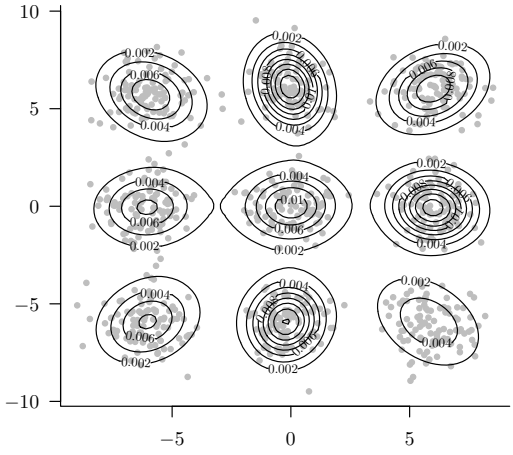
7 Acknowledgements

The authors are grateful to an AE and three anonymous Referees for insightful comments and suggestions. P. De Blasi and I. Prünster are supported by MIUR, PRIN Project 2015SNS29B. The work of A.F. Martínez is supported by SE-SES-DGESU-PRODEP grant UAM-PTC-697 and by CONACyT grant 241195. R.H. Mena is supported by CONACyT grant 241195. The work was concluded during a visit by R.H. Mena to the Department of Statistics & Data Sciences at the University of Texas at Austin. Hospitality from the department is gratefully acknowledged as the support from a Fulbright Scholarship.

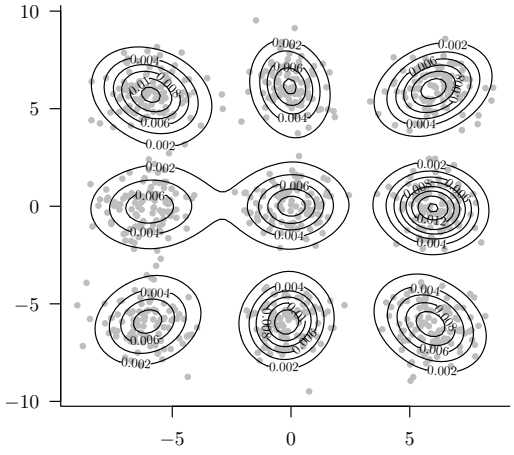


(a) Case $s = 1$

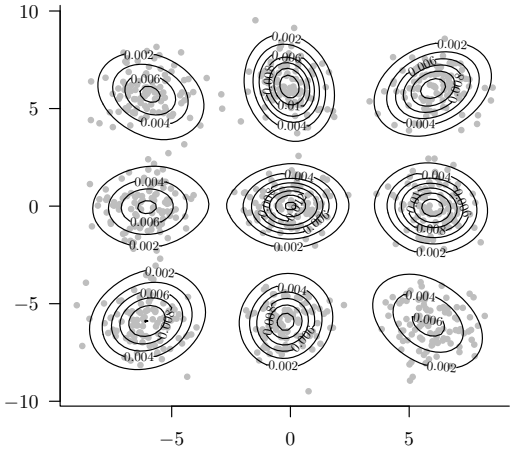
(b) Case $s = 2$



(c) Case $s = 5$



(d) Case $s = 10$



(e) Case $s = 25$

Figure 7: Contour plots for the estimated density of model (15) using negative binomial weights for different values of s and hyperparameters fixed such that $E(K_n) = 3$ *a priori*.

A Simulation of truncated distributions

In this section, we provide a sampling scheme to simulate from left-truncated Poisson and negative binomial distributions.

A.1 Truncated Poisson distribution

Suppose a random number x from a truncated Poisson distribution is required. A random variable $X \sim \text{Poi}(\lambda)$, with $\lambda > 0$, has probability mass function given by

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda} \mathbf{1}(x \geq 0).$$

If truncation is from the right, namely there is a $\tau \in \mathbb{Z}^+$ such that $\Pr(X > \tau) = 0$, simulation from it is straightforward since the resulting density has a finite support $\{0, 1, \ldots, \tau\}$. On the other hand, truncation from the left presents some issues, but these can be resolved via data augmentation, which can be done by means of a Gibbs sampler.

Suppose X is a random variable distributed Poisson with parameter λ and left-truncated at $\tau \in \mathbb{Z}^+$, that is, $\Pr(X < \tau) = 0$. Using a Gibbs sampler, it is required to sample from a density proportional to $\lambda^x / x! \mathbf{1}(x \geq \tau)$. Notice first that

$$\frac{1}{x!} = \frac{1}{(x-\tau)!} \frac{(x-\tau)!}{x!} = \frac{1}{(x-\tau)!} \frac{\Gamma(x-\tau+1)}{\Gamma(x+1)} \frac{\Gamma(\tau)}{\Gamma(\tau)} = \frac{\beta(x-\tau+1, \tau)}{\Gamma(\tau)(x-\tau)!}, \quad (16)$$

where $\beta(a, b) = \int_0^1 z^{a-1}(1-z)^{b-1} \mathrm{d}z$ is the beta function. Using data augmentation with $z \sim \text{beta}(x - \tau + 1, \tau)$, the sampler has to simulate from the bivariate density

$$p(x, z) \propto \frac{\lambda^x}{(x-\tau)!} z^{x-\tau} (1-z)^{\tau-1} \mathbf{1}(x \geq \tau).$$

From this, the full conditional distribution for x is given by

$$p(x|z) \propto \frac{(\lambda z)^{x-\tau}}{(x-\tau)!} \mathbf{1}(x \geq \tau),$$

which corresponds to a Poisson distribution with parameter λz shifted at τ. Therefore, a Gibbs sampler draws values at iteration t as follows:

$$z^{(t)} \sim \text{Be}(x^{(t-1)} - \tau + 1, \tau),$$
$$x^{(t)} \sim \text{Poi}(\lambda z^{(t)}) + \tau,$$

for an initial value $x^{(0)}$ and given λ and τ.

A.2 Truncated negative binomial distribution

In a similar way to the truncated Poisson distribution, simulating from a right-truncated distribution is straightforward, but not when it is left-truncated. The negative binomial probability mass function is given by

$$f(x) = \binom{x+r-1}{x} p^r (1-p)^x \mathbf{1}(x \geq 0),$$

for $r > 0$ and $0 < p < 1$.

Suppose that it is required to sample from a negative binomial distribution left-truncated at $\tau \in \mathbb{Z}^+$. This can be also done using a Gibbs sampler. In order to do it, notice that using (16),

$$p(x) \propto \frac{\beta(x-\tau+1,\tau)\Gamma(x+r)}{(x-\tau)!}(1-p)^x \mathbf{1}(x \geq \tau).$$

If the beta and gamma functions are substituted by their integral expressions, the resulting density is

$$p(x,z,v) \propto \frac{(1-p)^x}{(x-\tau)!} z^{x-\tau}(1-z)^{\tau-1} v^{x+r-1} e^{-v} \mathbf{1}(x \geq \tau),$$

and the full conditional distribution of x is

$$p(x|z,v) \propto \frac{((1-p)vz)^{x-\tau}}{(x-\tau)!} e^{-(1-p)vz} \mathbf{1}(x \geq \tau),$$

which is a Poisson distribution with parameter $(1-p)vz$ and shifted at τ. Therefore, using data augmentation, with $z \sim \text{beta}(x-\tau+1,\tau)$ and $v \sim \text{gamma}(x+r,1)$, a Gibbs sampler draws values at iteration t as follows:

$$\begin{aligned} z^{(t)} &\sim \text{Be}(x^{(t-1)}-\tau+1,\tau), \\ v^{(t)} &\sim \text{Ga}(x^{(t-1)}+r,1), \\ x^{(t)} &\sim \text{Poi}((1-p)z^{(t)}v^{(t)})+\tau, \end{aligned}$$

for an initial value $x^{(0)}$, and where r, p and τ are given values.

To the best of our knowledge, there is only one alternative procedure for sampling from these left-truncated distributions [Geyer, 2007], based on rejection sampling, but the acceptance ratio is low for some parameter values.

References

N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular Variation*. Cambridge University Press, 1987.

P. G. Bissiri and A. Ongaro. On the topological support of species sampling priors. *Electronic Journal of Statistics*, 8(1):861–882, 2014.

F. Camerlenghi, A. Lijoi, P. Orbanz, and I. Prünster. Distribution theory for hierarchical processes. *The Annals of Statistics*, 47(1):67–92, 2019.

P. De Blasi, A. Lijoi, and I. Prünster. An asymptotic analysis of a class of discrete nonparametric priors. *Statistica Sinica*, 23(3):1299–1321, 2013.

P. De Blasi, S. Favaro, A. Lijoi, R. H. Mena, I. Prünster, and M. Ruggiero. Are Gibbs-Type Priors the Most Natural Generalization of the Dirichlet Process? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):212–219, 2015.

M. D. Escobar and M. West. Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.

W. Feller. On a General Class of "Contagious" Distributions. *Annals of Mathematical Statistics*, 14(4):389–400, 1943.

T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

S. Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Springer, 2006.

R. Fuentes-García, R. H. Mena, and S. G. Walker. A nonparametric dependent process for Bayesian regression. *Statistics & Probability Letters*, 79(8):1112–1119, 2009.

R. Fuentes-García, R. H. Mena, and S. G. Walker. A New Bayesian Nonparametric Mixture Model. *Communications in Statistics - Simulation and Computation*, 39(4):669–682, 2010.

C. J. Geyer. Lower-Truncated Poisson and Negative Binomial Distributions, 2007.

S. Ghosal and A. van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press, 2017.

A. Gnedin, J. Pitman, and M. Yor. Asymptotic laws for regenerative compositions: gamma subordinators and the like. *Probability Theory and Related Fields*, 135(4):576–602, Aug 2006. ISSN 1432-2064. doi: 10.1007/s00440-005-0473-0.

J. E. Griffin and M. F. J. Steel. Stick-breaking autoregressive processes. *Journal of Econometrics*, 162:383–396, 2011.

L. Gutiérrez and F. A. Quintana. Multivariate Bayesian semiparametric models for authentication of food and beverages. *The Annals of Applied Statistics*, 5(4):2385–2402, 2011.

L. Gutiérrez, E. Gutiérrez-Peña, and R. H. Mena. Bayesian nonparametric classification for spectroscopy data. *Computational Statistics & Data Analysis*, 78:56–68, 2014.

S. Hatjispyros, C. Merkatas, T. Nicoleris, and S. Walker. Dependent mixtures of geometric weights priors. *Computational Statistics and Data Analysis*, 119:1–18, 2018.

M. Kalli, J. E. Griffin, and S. G. Walker. Slice sampling mixture models. *Statistics and Computing*, 21(1):93–105, 2011.

S. Karlin. Central limit theorems for certain infinite urn schemes. *J. Math. Mech*, 17(24): 373–401, 1967.

J. Kingman. *Poisson Processes*. Clarendon Press, 1993.

R. Korwar and M. Hollander. Contribution to the theory of Dirichlet processes. *Ann. Probab.*, 1:705–711, 1973.

A. Lijoi, R. H. Mena, and I. Prünster. Bayesian Nonparametric Estimation of the Probability of Discovering New Species. *Biometrika*, 94(4):769–786, 2007a.

A. Lijoi, R. H. Mena, and I. Prünster. Controlling the reinforcement in bayesian nonparametric mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):715–740, 2007b.

Lijoi, Antonio and Nipoti, Bernardo and Prünster, Igor. Dependent mixture models: Clustering and borrowing information. *Computational Statistics & Data Analysis*, 71:417–433, 2014.

A. Y. Lo. On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics*, 12(1):351–357, 1984.

G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley-Interscience, 2000.

R. H. Mena and S. G. Walker. On the Bayesian Mixture Model and Identifiability. *Journal of Computational and Graphical Statistics*, 24(4):1155–1169, 2015.

R. H. Mena, M. Ruggiero, and S. G. Walker. Geometric stick-breaking processes for continuous-time Bayesian nonparametric modeling. *Journal of Statistical Planning and Inference*, 141(9):3217–3230, 2011.

X. Nguyen. Inference of global clusters from locally distributed data. *Bayesian Analysis*, 5 (4):817–845, 12 2010.

O. Papaspiliopoulos and G. O. Roberts. Retrospective markov chain monte carlo methods for dirichlet process hierarchical models. *Biometrika*, 95(1):169–186, 2008.

K. Pearson. Contributions to the Mathematical Theory of Evolution. *Philosophical Transactions of the Royal Society of London A*, 185:71–110, 1894.

J. J. Quinlan, G. L. Page, and F. A. Quintana. Density regression using repulsive distributions. *Journal of Statistical Computation and Simulation*, 88(15):2931–2947, 2018.

S. Richardson and P. Green. On Bayesian Analysis of Mixtures with an Unknown Number of Components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59:731–792, 1997.

B. Scarpa and D. B. Dunson. Enriched stick-breaking processes for functional data. *J. Amer. Statist. Assoc.*, 109:647–660, 2014.

Y. Shi, M. Martens, A. Banerjee, and P. Laud. Low Information Omnibus (LIO) Priors for Dirichlet Process Mixture Models. *Bayesian Analysis*, 14(3):677–702, 2019.

D. Titterington, A. Smith, and U. Makov. *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York, 1985.

S. Wade, S. G. Walker, and S. Petrone. A Predictive Study of Dirichlet Process Mixture Models for Curve Fitting. *Scandinavian Journal of Statistics*, 41(3):580–605, 2014.